Astronomical Data Analysis Software and Systems: XXIV
ASP Conference Series, Vol. 495
A. R. Taylor and E. Rosolowsky, eds.

Rule-based Cross-matching of Very Large Catalogs

Patrick M. Ogle, Joe Mazzarella, Rick Ebert, Dario Fadda, Tak Lo,
Scott Terek, Marion Schmitz, and the NED Team

Infrared Processing and Analysis Center, California Institute of Technology,
1200 East California Boulevard, Pasadena, CA 91125, USA

Abstract. The NASA Extragalactic Database (NED) has deployed a new rule-based cross-matching algorithm called Match Expert (MatchEx), capable of cross-matching very large catalogs (VLCs) with > 10 million objects. MatchEx goes beyond traditional position-based cross-matching algorithms by using other available data together with expert logic to determine which candidate match is the best. Furthermore, the local background density of sources is used to determine and minimize the false-positive match rate and to estimate match completeness. The logical outcome and statistical probability of each match decision is stored in the database and may be used to tune the algorithm and adjust match parameter thresholds. For our first production run, we cross-matched the GALEX All Sky Survey Catalog (GASC), containing nearly 40 million NUV-detected sources, against a directory of 180 million objects in NED. Candidate matches were identified for each GASC source within a 7.″5 radius. These candidates were filtered on position-based matching probability and on other criteria including object type and object name. We estimate a match completeness of 97.6% and a match accuracy of 99.75%. Over the next year, we will be cross-matching over 2 billion catalog sources to NED, including the Spitzer Source List, the 2MASS point-source catalog, AllWISE, and SDSS DR 10. We expect to add new capabilities to filter candidate matches based on photometry, redshifts, and refined object classifications. We will also extend MatchEx to handle more heterogenous datasets federated from smaller catalogs through NED's literature pipeline.

1. Introduction

1.1. Background

The NASA Extragalactic Database (NED) provides a comprehensive fusion of multiwavelength data on extragalactic astrophysical objects, from data published in the astronomical literature and large, online catalogs (Helou & Madore 1988; Helou 1990; Mazzarella et al. 2007). With a rapid increase in data volume from space and ground-based surveys, NED is developing new methods for keeping apace. Human-interactive matching of catalog sources to NED objects is impractical for very large catalogs (VLCs) with $> 1 \times 10^7$ sources. Machine matching is faster, more accurate, and enables consistent application of rules. All catalog source and NED object attributes, continuous (e.g., position) or discrete (e.g., object type), are potential match discriminants. Machine matching rules systematize and codify decades of NED experience. Match metrics and outcome are tabulated in the database for statistical analysis. We chose the 40 million source GALEX All Sky Survey Catalog as the first VLC for

NED to tackle with our new rule-based machine matching algorithm. NED will continue to ingest and integrate even larger VLCs over the next few years, including the the Spitzer Source List, the 2MASS point source survey, AllWISE, and SDSS DR10.

Cross-matching is a central component of the Virtual Observatory concept, because it is a prerequisite for combining multiwavelength data (Malkov, O. et al. 2012). One important application is the identification of rare objects of interest by their SEDs, for example brown dwarfs selected by their SDSS-2MASS colors (Metchev et al. 2008; Geissler et al. 2011). Methods for cross-matching VLCs typically involve selecting candidate matches between two catalogs within a pre-defined separation threshold. Photometry may be used to exclude unlikely matches or to identify matches that meet color criteria for the object class of interest. Other algorithms utilize Bayesian statistics to select the most likely match, based on any number of parameters (Budavari & Szalay 2008). We have opted for an approach that begins with positional matching then applies additional criteria to select among match candidates, making use of the rich array of parameters available in NED. We measure the local density of background objects in the vicinity of each source to estimate the Poisson likelihood that an object is either a good match or a background object.

For NED, we make a distinction between entries in an incoming catalog (catalog sources) and the unique entries in the NED object directory that we match them to (NED objects). Catalog sources are typically listed as detections at one or more wavelength bands and have a unique catalog designation. NED objects are intended to represent unique astrophysical objects. For each object, NED provides cross-identifications to any catalog sources that NED has cross-matched to them and any associated positions, redshifts, photometric data, diameter measurements, classifications, morphologies, or other descriptors.

2. Methodology

2.1. Cross-matching VLCs with NED

In order to cross match a catalog against NED, we take the following steps. First, we load catalog source data (position, name, detection wavelength, photometry) into the database before matching. At this point the names and positions of VLC sources with positions may be made immediately available for perusal in NED. Next, we perform a positional search for match candidates in the database with Cone Search (CSearch). Then we run Match Expert (MatchEx) on a representative sample of match candidates to tune match rules and thresholds. After statistical optimization, we run MatchEx on the entire VLC. Any matches, new objects, or associations are loaded into the database by the ObjectLoader.

2.2. Match Candidate Selection

We use the PostgreSQL stored procedure CSearch to select all NED objects within a fixed search radius R_s of each catalog source as match candidates. The number of background sources is counted within a fixed background radius R_b, for use in computing the Poisson match probability. We also search for neighboring catalog sources within R_s of each catalog source, in order to identify candidate match conflicts.

2.3. Matching with MatchExpert

We use the python program MatchEx to select the best match candidate (if any) to each catalog source. MatchEx operates on source and object parameters from CSearch output. MatchEx currently uses source and object positions, position uncertainties, separation s, types (e.g., UvS, QSO), names, background object density n, and telescope beam size. The separation uncertainty σ is taken to be the sum in quadrature of the catalog and NED position uncertainties. MatchEx uses conditional logic to determine which NED objects in the search region centered on each catalog source are acceptable matches. The match criteria include thresholds on separation s, normalized separation $r = s/\sigma$, Poisson probability P, and type and name preferences or exclusions. Additional criteria for future VLC matching may utilize photometry, redshifts, and detailed morphological or spectral classifications.

For any given catalog source there are three possible MatchEx outcomes. If one object meets all match criteria then make the cross-ID. If no single object meets all match criteria then create a new NED object and any associations. If multiple (N) sources match a single object, then create N new objects and $(N + 1)N/2$ associations.

2.4. Position-based Match Probability

MatchEx puts position-based probabilistic matching on a firm statistical basis. The Poisson statistic is used to estimate the false-positive rate for pure position matching. However, this value should be regarded as an upper limit since the MatchEx selection algorithm uses additional source and object parameters to eliminate background and improve match accuracy.

The number N, and mean local surface density $n = N/(\pi R_b^2)$ of NED objects is measured within the background radius R_b, and used to estimate the background contamination rate from the Poisson match probability. The Poisson probability is computed from the Poisson distribution $P_s(x = k) = \langle N_s \rangle^k \exp(-\langle N_s \rangle)/k!$, where x is the number of sources found within separation s, and $\langle N_s \rangle = n\pi s^2 = N(s/R_b)^2$ is the expected number of background objects within s. For each source-object match candidate, we compute the likelihood that $k = 0$ background sources are found closer to the source than s, $P = P_s(x = 0) = \exp(-\langle n \rangle)$, giving $P = exp(-N(s/R_b)^2)$. Summing up the Poisson probabilities for all matches gives the false-positive match rate $f_p = (1 - \sum P)/N_G/100$. Note that the f_p value has to be determined experimentally from the MatchEx matching results. We can tune the false-positive match rate by raising or lowering the Poision probability match threshold P_t.

For the most efficient search, the ratio R_b/R_s =(background radius)/(search radius) needs to be adjusted to the Poisson probability threshold. If this ratio is too small, objects found in an outer annulus of the search region will have Poisson probabilities below the threshold, even when there are zero additional objects inside the background radius. This ratio is optimal for $N = 1$ background sources when $P = \exp(-1 \times (R_s/R_b)^2) = P_t$. For example, $R_b/R_s = 3.09$ is the most efficient value to use for a Poisson threshold of 0.90.

2.5. Match Selection Logic

The match selection logic used by MatchEx is illustrated in Figure 1 and summarized as follows. A match is made between a catalog source and a NED object when there is a single object in the search radius with $r = s/\sigma \leq r_t$, $P \geq P_t$, $s \leq R_s$, of allowed object

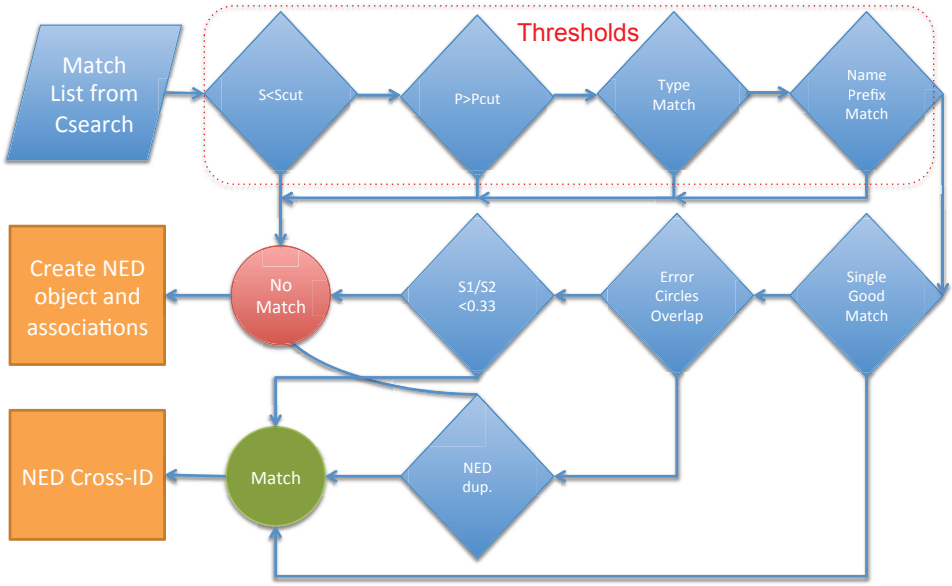

Figure 1. Match selection logic.

type, and that does not overlap another NED object. Overlap occurs when two or more NED objects have overlapping position error circles (95% enclosed probability), such that there is a significant chance that they represent the same astrophysical object. If there are multiple objects matching the above criteria, and only one of them has a preferred object type or preferred cross-ID prefix, then that object is a match. Otherwise, if there are multiple objects matching the above criteria, a match is made to the closest object if the ratio of separation to the second closest object is $s_1/s_2 < 0.33$. However, if there are two or more objects of the same type within $s \leq 1.0''$, these are assumed to be duplicate NED objects, and the source will be matched to the object with the most cross-ID's (a measure of popularity).

2.6. Preferred and Excluded Object Types and Names

A good way to reduce the effective NED background object confusion is to exclude illogical, problematic, or unlikely object types or names, or to preferentially match preferred (likely) object types or names. For an optical catalog like GALEX ASC, objects with the following preferred NED object types should be excluded as matches: AbLS, GClstr, GGroup, GammaS, XrayS, SmmS, RadioS, Nova, and SN. The AbLS, GClstr, and GGroup types are excluded because they should not match to a (single) optical source. A generic GammaS, X-rays, SmmS, or RadioS that has not previously been cross-ID'd with an optical source will typically have a large beam and large positional uncertainty not suitable for matching. Matches to variable sources such as Novae and SNe could be allowed if constraints on observation date were considered.

2.7. Associations

Two types of associations ($\rightarrow$,$\leftrightarrow$) may be created between catalog sources and NED objects: ErrorOverlap and InBeam (defined below). In general, a NED association

record captures a relationship between two different NED objects. An association may either be symmetric ($a \leftrightarrow b$), or asymmetric ($a \rightarrow b$). While associations are indicated in the MatchEx results table wherever they occur, they are added to a NED association table only when a catalog source is not matched to the corresponding NED object. The first type of association, ErrorOverlap, is created where a catalog source has a position error circle that overlaps that of a NED object and vice versa. This is an example of a symmetric association ($\leftrightarrow$). The second type of association, InBeam, indicates a NED object that falls inside a catalog source beam ($s < 4.8''$ for GALEX). This is an asymmetric association ($\rightarrow$), useful for indicating when a catalog source may combine photons from multiple NED objects, or if a candidate match was rejected for other reasons, even though it was the only NED object that fell inside the source beam.

3. Input Data

The GALEX All-Sky Survey Source Catalog (GASC, 2012GASC..C....0000S, `http://www.galex.caltech.edu`) contains 39,570,031 NUV-selected sources, corresponding to $> 3\sigma$ detections in the NUV band. The GASC survey covers 26,300 square degrees (8.0 sr, or 63.8% of the sky), consisting of all GALEX exposures with exposure times < 800 sec (typically 100 sec), imaged to a mean depth of NUV= 20.5 mag (AB). Gaps in sky coverage include the Galactic plane, Magellanic Clouds, and regions containing bright stars. The GALEX imaging FWHM is $5.3''$ in the NUV band, giving a Gaussian beam diameter of $9\farcs6$ at 10% peak flux.

GASC Photometry consists of measurements in the NUV (λ = 2316Å, $\Delta\lambda$ = 1000Å) and FUV (λ = 1539Å, $\Delta\lambda$ = 400Å) bands, using a number of methods and a range of apertures. NED has selected photometry from two different methods for each of the FUV and NUV bands to include in its photometric database. The first method gives the Kron flux in an elliptical aperture, which is appropriate for extended sources. The second method gives the flux in a $7.5''$ diameter aperture.

GASC was matched to candidates selected from the NED 23.7 production database in 2013 October. That version of NED contained roughly 180 million unique objects derived from 90, 211 references, with 222 million multiwavelength cross-IDs.

4. Tuning MatchEx

In order to tune the MatchEx algorithm, match thresholds, and performance, a series of test runs were conducted on subsets of the full GASC 40M catalog. These subsets were contained in a circular region centered at (RA,Dec)=(200,+30) in the middle of the Sloan Digital Sky Survey (SDDS) North Galactic Pole (NGP) survey region; and outside the SDSS survey region, centered at (RA,Dec)=(60,+30). We used an on-SDSS test with 10^5 GASC sources (GASC 100K) as the primary means to characterize and tune the matching algorithm and thresholds.

The search radius was initially selected to be twice the GALEX beamsize ($9\farcs6$). It was decreased to to $R_s = 7.5''$ after considering the observed separation distribution of GASC-NED matches. Note that CSearch actually delivers match candidates out to $2R_s = 15''$ in order to catch any potential match conflicts where more than one GASC source matches to the same NED object. A background radius of $R_b = 6.2R_s = 46\farcs5$

was used for measuring the local background density for use in calculating the Poisson match likelihood.

The goal in tuning the match thresholds is to minimize both the false positive f_p and false negative f_n match rates. These two rates tend to offset one another, with more strict thresholds reducing the false positive rate at the expense of increased false negative rate. Where this trade-off is optimized depends on how much weight is given to accuracy versus completeness. Because we think it is much worse to make an incorrect match than to miss a potentially good match, we use the combined error rate $f_e = f_n + 10 \times f_p$ as our performance metric for MatchEx.

The Poisson statistic is the primary statistic that we use to determine match likelihood, and it is directly related to the false positive match rate. We ran the GASC 100K test several times with Poisson statistic threshold in the range $P_t = 0.82 - 0.98$ to find the dependence of f_p, f_n, and f_e on P_t (Figure 3). We find that f_e is minimized for a Poisson threshold of $P_t = 0.90$. The minimum in f_e is rather broad, so the precise value of P_t does not make much difference in the range $P_t = 0.88 - 0.92$. We have not yet made a detailed study of the impact of other thresholds and selection criteria (search radius, object type exclusions, and object name preferences) on the match error rate.

We also used a threshold of $r = s/\sigma = 3.5$, with the aim of limiting the match incompleteness outside of the search area to 0.2%. However, the non-Gaussian distribution of separation errors, with extra matches found at large separations led to greater incompleteness.

Finally, GASC source matches to NED objects with object types or cross-ID types of UvES (UV-excess) or with object name or cross-ID name beginning with the string "GALEX" were preferred. This means that if there was only one UvES or GALEX match candidate within the search radius and it fell within the match thresholds for all other parameters, it was automatically selected as the best match.

5. Results

5.1. Overall Statistics

From the 39,570,031 GASC UV sources and 23,301,552 NED object match candidates, there are 10,595,382 (26.8%) matches to NED objects and 28,974,649 (73.2%) no matches, of which 26,984,670 (68.2%) are no-matches in NED blank fields. The remaining 1,992,979 (5.0%) of no-matches occur in non-blank fields, including fields with one match candidate (1,027,515= 2.6%), two match candidates (471,300= 1.2%), 3 match candidates (260,517= 0.7%), and 4 or more match candidates (233,647= 0.6%).

We present distributions for the position-based matching parameters in Figure 4 for MatchEx selected matches, compared to the unfiltered CSearch selected match candidates. The number of match candidates is expected to increase linearly with separation at large separation ($s > 4''$), for uniform mean background density: $N = \langle n \rangle \pi s^2$, $dN/ds = 2\pi \langle n \rangle s$. The observed background increases faster than linearly at separations $s > 12''$, for an unknown reason. We fit the background at $s = 6 - 12''$ and find a slope of 5.33×10^3 arcsec^{-1}/(0.05″ bin)= 1.07×10^5 arcsec^{-2}. This corresponds to a mean background density of $\langle n \rangle = 1.07 \times 10^5$ arcsec$^{-2}/(2\pi) = 1.70 \times 10^4$ arcsec^{-2}.

By integrating under the linear background fit, we estimate the number of background objects within $s_t = 7.5''$ to be 2,980,000 (21.5% of match candidates). This is

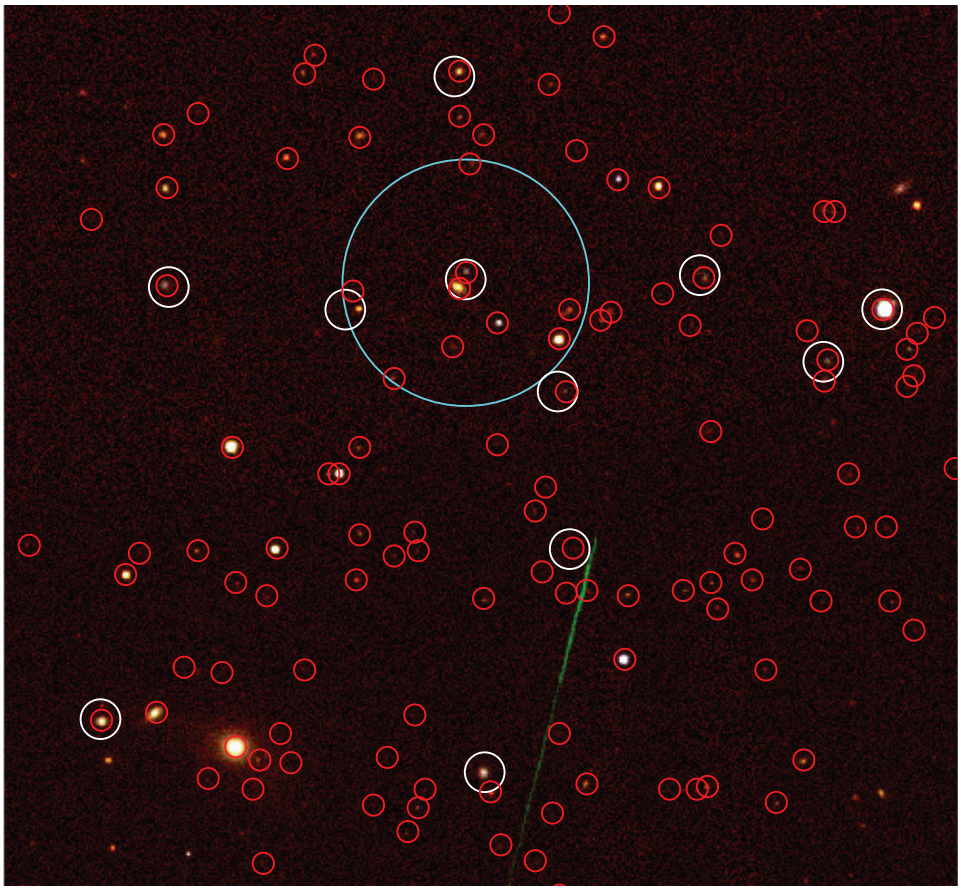

Figure 2. Overlay of 7.″5 radius search regions for GALEX ASC NUV sources (white circles) on $6' \times 6'$ SDSS DR6 gri image, centered at (RA,Dec)=(200,+30). One of the regions used to estimate the local background object density surrounding one of the sources is shown in cyan. The locations of NED objects are indicated by red circles.

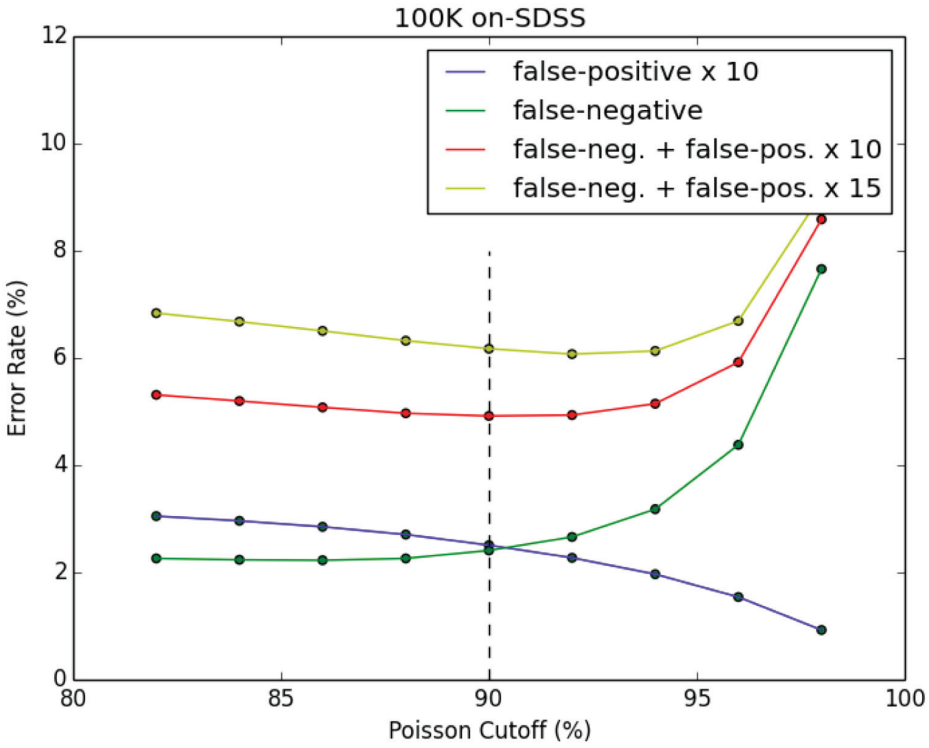

Figure 3. Optimization of match error metric vs. Poisson statistic threshold. The false positive rate drops, while the false negative rate rises with Poisson threshold P_t. The combined match error metric $f_e = f_n + 10 \times f_p$ has a minimum at $P_t = 0.90$.

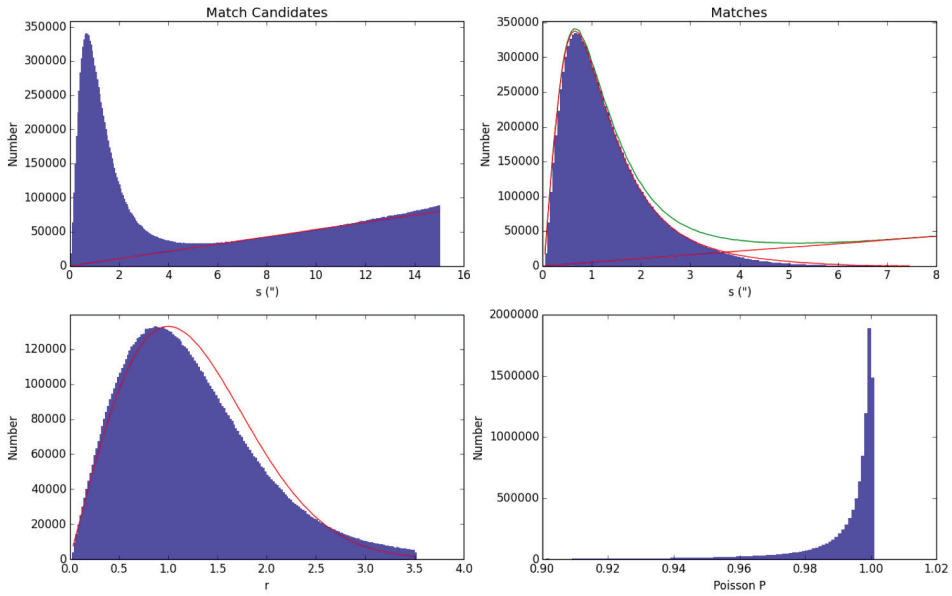

Figure 4. Distribution of separation s, normalized separation $r = s/\sigma$, and Poisson match statistic P for GASC-NED matches The linear background estimate is indicated by the red line in the upper-left panel. The upper-right panel compares the distribution of match separations to candidate matches (green line) and background-subtracted candidate matches (red line). The normalized separation histogram is compared to the derivative of the Gaussian error distribution with $\sigma_r = 0.9, 1.0$ (red curves in the lower left panel). There is significant deviation from a Gaussian distribution of positional errors, perhaps indicating a range in positional errorbar under or overestimation.

similar to the number of rejected match candidates (3,275,249 or 23.6%) inside this radius. The actual match separation distribution is close to the one obtained by subtracting the linear background fit from the match candidate separation distribution, indicating that MatchEx does a good job of eliminating background NED sources as matches. The small difference between the two at $s = 3.5 - 7''$ gives an estimate of $f_n = 2.4\%$ for the false negative match rate. As mentioned previously, this is considerably greater than the Gaussian incompleteness estimate 0.2%, because the actual error distribution has an extended non-Gaussian tail.

The distribution of dN/dr vs. r is compared to the derivative of a Gaussian function (Fig. 2). The peak of the dN/dr vs. r distribution lands near 0.9σ, showing that the scale of the combined GASC and NED position errorbars may be overestimated by 10%. However, there is an excess of matches at $r > 1.5\sigma$ compared to a Gaussian with standard deviation $\sigma_r = 0.9\sigma$. This shows that the error-distribution in separation is not perfectly matched to a Gaussian distribution

The Poisson probability density distribution for CSearch match candidates peaks sharply at a value of $P = 1$. There is a tail of match candidates with probabilities $0 < P < 0.9$, corresponding to background NED objects. The chosen Poisson match threshold of $P_t = 0.9$ roughly matches the location in the distribution where the density of true matches begins to exceed the number of false-positives.

5.2. Comparison of GASC and SDSS Photometry

After GASC-NED matches and GASC photometry are loaded into NED, the corresponding spectral energy distributions are automatically populated with the new GASC photometry.

5.3. Object Type and Catalog Prefix Statistics

Table 1 gives the breakdown of GASC-NED candidate matches and matches by NED preferred object type (60 in all). The top-12 most abundant types account for 99.82% of the total match candidates. There are 14 types not represented by match candidates. The 3 types RadioS, XrayS and GGroup were excluded as matches by MatchEx for reasons explained previously. Therefore, the top-9 most abundant types account for 99.94% of the total matches. The remaining 0.06% of matches are distributed amongst 34 less-common NED object types.

We focus on matches to NED objects in the top-9 matched object type categories. Galaxies (G) are most abundant, making up roughly 61% of GASC-NED matches. Stars (*) are second, making up roughly 36% of matches. The remaining 3% of matches are to NED object types VisS, QSO, !*, UvES, IrS, WD*, and GPair in order of abundance. To make sense of the object type abundance patterns for matches, we take the ratio of their abundances divided by the overall abundances in NED (last column of Table 1). Interestingly, matches to objects of type QSO, !*, UvES, and WD are significantly enhanced, by factors of 9-14 relative to their NED abundances, which may reflect the tendency for these types of objects to be relatively UV-bright. On the other hand, matches of type IrS are under abundant, with a ratio of 0.32 with respect to NED.

The top-11 catalog prefix names account for 98.1% of GASC-NED matches. The top-4 preferred catalog prefix names (SDSS, APMUKS(BJ), MRSS, and 2MASX) account for 95.8%. NED objects with SDSS preferred names constitute the large majority (75.9%) of GASC-NED matches. There were 57,633 matches to objects with GALEX preferred names.

6. Future Improvements

Match accuracy may be improved by considering the full array of object data compiled by NED, including redshifts, photometry, diameters, and detailed classifications. Spectroscopic redshifts are only available for a small fraction of NED objects, but can provide a strong constraint for matching catalog sources with redshifts. Photometric constraints will be applicable to most of the VLCs that NED will be matching. However, such constraints should be relatively weak and applied to the tail of the color distribution (e.g., $|NUV - g| < 10$ mag for GALEX) since we do not want to bias matches against objects with unusual SEDs. We also plan to make use of object size, overlap, and type for extended sources. A small, offcenter point source within a large extended source is likely to be a part of the extended source rather than a match to the extended source. Parts of galaxy types such as HII regions, star clusters, and variable stars should not be matched to the full galaxy.

Acknowledgments. The NASA/IPAC Extragalactic Database (NED) is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

Table 1. Match Counts by Object Type

Type Code	NED	Cand.	Match	Matched %	NED %	Cand. %	Match %	Abund. Ratio
G	102,386,926	13,816,821	6,447,394	46.6	57.2	59.3	60.8	1.06
*	68,268,375	8,391,557	3,801,913	45.3	38.1	36.0	35.9	0.94
VisS	1,821,299	321,549	99,406	30.9	1.02	1.38	0.94	0.92
IrS	1,661,077	239,803	31,370	13.1	0.93	1.03	0.30	0.32
RadioS	2,025,701	152,692	0	0.0	1.13	0.66	0	0
QSO	163,260	98,648	86,681	87.9	0.091	0.42	0.82	9.01
!*	89,416	69,301	58,859	84.9	0.050	0.30	0.56	11.2
UvES	129,330	62,752	56,238	89.6	0.072	0.27	0.53	13.6
XrayS	407,843	52,820	0	0.0	0.23	0.23	0	0
GGroup	92,910	34,739	0	0.0	0.052	0.15	0	0
GPair	26,669	9,759	2,312	23.7	0.015	0.042	0.022	1.47
WD*	9,461	7,951	7,574	95.3	0.0053	0.034	0.071	13.4

References

Bonnarel, F. et al. 2000, ASPC, 216, 239
Budavari, T. & Szalay, A. S. 2008, ApJ, 679
Geissler, K., Metchev, S., Kirkpatrick, J. D., Berriman, G. B. & Looper, D. 2011, ApJ, 732, 56
Gezari et al. 2013, ApJ, 766, 60
Helou, G., Madore, B. F., Schmitz, M., Wu, X., Corwin, H. G., Jr., Lague, C., Bennett, J., & Sun, H. 1995, ASSL, 203, 95; in "Information & on-line data in astronomy"
Helou, G. 1990, BICDS, 38, 7
Helou, G. & Madore, B. 1988, ESOC, 28, 335
Mazzarella, J. M. & the NED Team. 2007, ASPC, 376, 153
Malkov, O., Dluzhnevskaya, O., Karpov, S., Kilpio, E., Kniazev, A., Mironov, A., & Sichevskij, S. 2012, BaltA, 21, 319
Metchev, S. A., Kirkpatrick, J. D., Berriman, G. B, & Looper, D. 2008, ApJ, 676, 1281